Developing speech input for virtual reality applications: A reality based interaction approach

Alex W. Stedmon*, Harshada Patel, Sarah C. Sharples, John R. Wilson

Virtual Reality Applications Research Team (VIRART), University of Nottingham, Nottingham, UK

Received 15 January 2010; received in revised form 20 September 2010; accepted 27 September 2010
Communicated by E. Motta
Available online 20 October 2010

Abstract

An input device should be natural and convenient for a user to transmit information to a computer, and should be designed from an understanding of the task to be performed and the interrelationship between the task and the device from the perspective of the user. In order to investigate the potential of speech input as a reality based interaction device, this paper presents the findings of a study that investigated speech input in a VR application. Two independent user trials were combined within the same experimental design to evaluate the commands that users employed when they used free speech in which they were not restricted to a specific vocabulary. The study also investigated when participants were told they were either talking to a machine (e.g. a speech recognition system) or instructing another person to complete a VR based task. Previous research has illustrated that when users are limited to a specific vocabulary, this can alter the interaction style employed. The findings from this research illustrate that the interaction style users employ are very different when they are told they are talking to a machine or another person. Using this knowledge, recommendations can be drawn for the development of speech input vocabularies for future VR applications.

© 2010 Elsevier Ltd. All rights reserved.

Keywords: Virtual reality; Speech input; Human factors; Reality based interaction

1. Reality based interaction for speech interfaces

An input device should be natural and convenient for a user to transmit information to a computer (Jacob et al., 1993), and should be designed from an understanding of the task to be performed and the interrelationship between the task and the device from the perspective of the user (Jacob et al., 1994). The design of an input device should match the perceptual needs of the user and, as a result, the integration of input devices should follow a user needs analysis in order to map their expectations onto the attributes of the overall virtual reality (VR) system (Kalawsky, 1996). Building on a sound understanding of user needs, it is important, therefore, to analyse the task in the correct level of detail, so that the VR system and the VE that is developed supports user interaction and overall application effectiveness (Stedmon, 2003). A VR input device should account for the type of manipulations a user has to perform and be designed, so that it adheres to natural mappings in the way that the device is manipulated, as well as permit movements that coincide with a user’s mental model of the type of movement required in a VE (Barfield et al., 1998). Despite such recommendations, there are very few established guidelines that detail what is required of an input device in terms of the parameters that pertain to a user and user performance within a VE (Kalawsky, 1996; Lee and Billinghurst, 2008).

With the development of reality based interfaces (RBIs) and new interaction styles that draw on users’ knowledge, experience and expectations of the real-world, the aim is to develop human–computer interaction (HCI) metaphors in a digital world that are more intuitive and less constrained by technology (Jacob et al., 2008). With this in mind Jacob et al. (2008) present a framework for RBIs based on four fundamental themes:

- Naïve physics: people have common sense knowledge about the physical world.
Body awareness and skills: people have an awareness of their own physical bodies and possess skills for controlling and coordinating their bodies.

Environment awareness and skills: people have a sense of their surroundings and possess skills for negotiating, manipulating and navigating within their environment.

Social awareness and skills: people are generally aware of others in their environment and have skills for interacting with them.

The themes play a prominent role in emerging interaction styles as they offer universal paradigms for interaction which are not bound by cultural differences (Jacob et al., 2008). In relation to speech input, the RBI framework provides a clear structure for exploring both human–human interaction (HHI) and human–machine interaction (HMI). This paper presents an overview of speech input as a reality based interaction device before presenting the findings of a study that investigated reality based interaction of using speech in VR applications. The findings are then discussed in relation to the RBI framework along with recommendations for speech interfaces.

2. Speech input as a reality based interaction device

Speech is one of the most natural forms of HHI which most of us are able to employ in our daily lives from an early age (Stedmon, 2003). It is still not fully understood how we learn the subtle rules of syntax and grammar but what is clear, however, is that speech is a familiar, convenient, and spontaneous part of the capabilities people bring to the situation of interacting with machines (Lea, 1980).

In general, the use of speech for HHI requires little or no specific training except perhaps for tasks requiring specialized domain knowledge and vocabulary. When conducting HMI speech may offer advantages over keyboard skills or manual input, which are difficult without practise and training (Stedmon, 2005). Speech input presents considerable opportunities to harness reality based interactions; however, there is often still a requirement to learn task grammar and develop suitable mental models that account for how a machine might respond to spoken inputs. There is considerable progress in developing multimodal interaction, such as gesture, video tracking, and electromagnetic sensing, illustrating that as much as 97% of commands involve some degree of gesture based communication and that 94% of speech commands are preceded by gesture commands (Lee and Billinghurst, 2008).

2.1. Speech input for VR applications

Speech offers the potential to liberate users and allow a greater degree of freedom to interact within VEs as traditional input devices are often unsuited to simultaneous 3D navigation and control (Stanger, 1997). Speech allows hands free operation and could be particularly useful for navigation through easily accessible menus and short cut commands (Stedmon, et al., 2003). Furthermore, multi-user recognition systems support speech input in collaborative VEs allowing many users to interact with each other and the VE at the same time.

There is an increasing variety of input devices on the market that have been designed for VR use, ranging from tradition mouse and joystick devices, to wands, data gloves, haptic feedback devices, and speech interfaces (D’Cruz, 1999). Such variety may lead to individuals selecting an inappropriate input device that could compromise overall task effectiveness as well as impact on their overall satisfaction in using the VR application (Stedmon, 2003).

3. A reality based investigation of speech input for VR applications

In order to investigate the potential of speech input as an RBI device for emerging VR applications, two user trials and a comparative analysis designed to investigate issues associated with free speech were conducted. Two independent user trials were combined within the same experimental design to evaluate the commands that users employed when they were not restricted to a specific vocabulary. The study investigated speech input when participants were told they were either talking to a speech recognition system or instructing another person to complete a VR based task. Previous research has illustrated that when users are limited to a specific vocabulary and under time pressure, they are likely to revert back to prelearnt or more intuitive vocabularies (Baber et al., 1996) or when under increased cognitive load, they will take longer to recall the appropriate vocabulary (Stedmon and Baber, 1999). Furthermore, the match between the language users employ when using a speech input system and the language that the system can accept, or the habitability of a system, is a key issue in system usability (Hone and Baber, 2001).

In order to achieve free speech without the technical limitations of current speech recognition systems, a ‘Wizard of Oz’ paradigm was employed in a similar fashion to Baber and Stammers (1989) and Lee and Billinghurst (2008). Participants believed they were using a speech recognition system but, unknown to them at the time, an experimenter manipulated the VE based on their spoken instructions. In this way, by allowing users to employ any commands they wished, it was possible to investigate the following:

- The variety of legal and illegal commands that were used to complete the task;
- The style of speech used when participants were told they were talking to a speech recognition system;
- The style of speech used when participants were told they were talking to another person.
3.1. Method

3.1.1. Participants
In each of the two independent user trials 12 participants were recruited. In the HMI Trial age ranged from 20 to 52 years; mean age = 31.5 years. In the HHI Trial age ranged from 21 to 53 years; mean age = 32.5 years. All participants were staff or students from the University of Nottingham with English as their first language and normal or corrected to normal vision. Most users (n = 21, 87.5%), assessed by informal interview, had not used VR applications or speech recognition systems in the past (n = 11, 91.7% in the HMI Trial; n = 10, 83.3% in the HMI Trial).

3.1.2. Apparatus
The VR system comprised of an 800 MHz laptop PC, running Superscape VRT software (Jeapes, 2007) with a data projector and a forward projection screen. User input was via a hand held microphone whilst the experimenter interacted with the VE using a joystick. The VE used for the trials was Netcard, which was developed by the University of Nottingham (D’Cruz, 1999) to assess the effectiveness of VR as a training tool for computer network card replacement (Fig. 1).

3.1.3. Design
The two independent user trials combined to form a between-subjects design. The independent variable was the mode of speech input which had two levels: human–machine interaction (HMI Trial) and human–human interaction (HHI Trial). The dependent variables for both the user trials included: task completion time; total number of commands used; number of words used; number of words per command; and the number of illegal commands in which a participant requested an action the VE could not support. Additional measures of the number of personal words (e.g. “can you move forward”); relative words (e.g. “put the screwdriver over there”), and polite words (e.g. “please open the door”, “put the screwdriver back, thank you”) that participants used were also analysed. As the focus of this study was on actual speech styles and task performance, no formal subjective measures were included although informal interview data were collected.

3.1.4. Procedure
Within the Netcard VR application, users were required to change a PC netcard using spoken commands. In order to investigate the concept of developing intuitive reality based speech commands, users were allowed to speak freely for navigation and object manipulation within the Netcard VE. To complete the task, users had to navigate to a desk with a computer on it; remove the computer cover; take out the old network card and replace it with a new card; and replace the cover. There were 22 successive stages in completing the task and users were given simple text based instructions for completing the task on an external overlay directly outside the VE. After completing the HMI trial, participants were informed that they had been speaking to the experimenter. Participants were paid for their time.

3.2. Results
The observations from both the user trials are described below and followed by a comparative analysis between the trials.

![Fig. 1. The Netcard VR application.](image)
3.2.1. Task completion time

Task completion time was calculated from when participants spoke their first command to when the final stage had been completed. In the HMI Trial, mean completion time = 35 min 18 s (SD = 15 min 17 s) with task completion times ranging from 18 min 42 s to 61 min 19 s. In the HHI Trial mean completion time = 30 min 33 s (SD = 8 min 20 s) with task completion times ranging from 19 min 48 s to 47 min 49 s.

3.2.2. Total number of commands

The total number of commands were analysed in each trial. In the HMI Trial, mean number of commands = 237.5 (SD = 87.16) ranging from 128 to 415 commands. In the HMI Trial, mean number of commands = 195.36 (SD = 83.37) ranging from 89 to 403 commands.

3.2.3. Number of different commands

The number of different commands that were used was calculated in the HMI Trial. In total, 348 different commands were used by all the participants in the HMI Trial to complete the task. Mean different commands = 55.5 (SD = 13.56) ranging from 37 to 72 commands. Whilst a high number of total commands were used, approximately 77% were the same for each participant.

3.2.4. Total number of words

The total number of words used to complete the task were analysed in each trial. In the HMI Trial, mean number of words = 500 (SD = 193.65) ranging from 269 to 980 words. In the HHI Trial, mean number of words = 675 (SD = 202.55) ranging from 398 to 1057 words.

3.2.5. Number of words per command

By dividing the total number of words used by the total number of commands, it was possible to calculate the mean number of words per command. In the HMI Trial, mean number of words per command = 2.16 (SD = 0.55) ranging from 1.74 to 3.62 words per command. In the HHI Trial, mean number of words per command = 3.73 (SD = 0.99) ranging from 1.92 to 4.91 words per command.

3.2.6. Number of computer prompts

When participants issued a command beyond the capability of the VE, they were requested by the computer to try a different command. In the HMI Trial, mean number of illegal commands = 12.46 (SD = 8.71) ranging from 2 to 29 prompts. In the HHI Trial mean number of illegal commands = 14.36 (SD = 7.38) ranging from 4 to 24 prompts.

3.3. Comparative analysis of user trials

All data were tested in order that they met the assumptions for parametric analysis and analysed using SPSS (v.15) statistical software using independent-samples T-tests. Mean and standard deviation data from the comparative analyses are presented in Table 1.

### Table 1

Mean and SD data for the HMI and HHI Trials.

<table>
<thead>
<tr>
<th>Task variable</th>
<th>HMI Trial</th>
<th>HHI Trial</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task completion time (min:s)</td>
<td>35:18 (15:17)</td>
<td>30:33 (8:20)</td>
<td>N.S.</td>
</tr>
<tr>
<td>Total no. of commands</td>
<td>237.5 (87.16)</td>
<td>195.36 (83.37)</td>
<td>N.S.</td>
</tr>
<tr>
<td>Total number of words</td>
<td>500 (193.65)</td>
<td>675 (202.55)</td>
<td>N.S.</td>
</tr>
<tr>
<td>No. of words per command</td>
<td>2.16 (0.55)</td>
<td>3.73 (0.99)</td>
<td>p &gt; 0.05</td>
</tr>
<tr>
<td>No. of computer prompts</td>
<td>12.46 (8.71)</td>
<td>14.36 (7.38)</td>
<td>N.S.</td>
</tr>
<tr>
<td>No. of personal words</td>
<td>0.42 (1.00)</td>
<td>11.75 (12.22)</td>
<td>p &gt; 0.01</td>
</tr>
<tr>
<td>No. of relative words</td>
<td>1.42 (2.35)</td>
<td>25.17 (22.49)</td>
<td>p &gt; 0.001</td>
</tr>
<tr>
<td>No. of polite words</td>
<td>0 (0)</td>
<td>1.58 (2.75)</td>
<td>N.S.</td>
</tr>
</tbody>
</table>

A significant effect was observed for the number of words per command \(t(22) = -2.124, p < 0.05\) (2-tailed), illustrating that participants in the HMI Trial issued shorter commands (mean = 2.16 words per command) than participants in the HHI Trial (mean = 3.73 words per command). A significant effect was also observed for the number of personal words used \(t(22) = -3.202, p < 0.01\) (2-tailed), illustrating that more personal words were used in the HHI Trial (mean = 11.75 words) than in the HMI Trial (mean = 0.42 words). In addition, a significant effect was observed for relative words used \(t(22) = -3.639, p < 0.001\) (2-tailed), illustrating that more relative words were used in the HHI Trial (mean = 25.17 words) than in the HMI Trial (mean = 1.42 words). No other significant effects were observed \(p > 0.05\).

4. Discussion

4.1. Comparing the HMI and HHI Trials

The overall results are summarised in Table 2. The findings from the independent user trials illustrate that there was a difference in the way participants talked when they had been instructed that they were talking to either a machine or another person. It was assumed that with participants using free speech in the HHI Trial, this would generate a large variety of commands. As the focus of this research was on reality based interaction styles for HMI, the number of different commands was analysed in greater detail to illustrate the variety of commands used in this context. The findings illustrate that participants repeated commands they had used previously in the trial and built up a reusable command structure as they progressed through the task. Participants used shorter commands and less relative terms when they were talking to a machine. From informal interviews, participants used short commands when talking to a machine because they thought the machine would not understand more complex commands. As a consequence, the commands were not detailed enough and participants had to repeat or elaborate them in order to complete the task. This resulted in participants taking longer and using more commands to complete the task in the HMI Trial than in the HHI Trial.
The notion that participants felt the machine might not understand complex commands, supports the concept of habitability between the vocabulary users employ when using a computer system and the command structure a system might recognise (Hone and Baber, 2001). When people address a machine rather than a human, they tend to use more explicit and direct questions, fewer pronouns, and a more restricted, formal, vocabulary (Richards and Underwood, 1984); short, succinct, and task specific commands (Graham et al., 1998) with up to 74% of speech commands being phrases of a few discrete words whilst only 26% are complete sentences (Lee and Billinghurst, 2008). These findings support the idea that users might treat speech recognition systems as an intelligent servant rather than a dialogue partner (Baber and Stammers, 1989).

In the HHI Trial the commands were longer and more complex (e.g. “go over to the table and pick up the screwdriver” or “undo 8th screw on the right”). However, the overall time to complete the task and the number of commands used were both lower in the HHI Trial when compared to the HMI Trial. The experimental and informal interview data for use of personal, relative, and polite words illustrated that participants assumed the person they were interacting with had some shared knowledge, viewpoint, or mental model of the VE and could relate the commands they heard to objects in the VE or activities they had previously conducted. It is interesting to note that whilst the observed differences were not significant, participants used more polite words when talking to another person. Conversely, in the HMI Trial, participants did not issue any polite words, reinforcing the idea that they believed they were talking to an inanimate interface.

By analysing the most and least common commands it is possible to begin to develop ideas regarding generic command structures. Movement in the VE was continuous (e.g. movement continued until another action was performed) and so the most common command in both user trials used was: stop. If movement in the VE was discrete (e.g. movement occurred incrementally at a specified distance each time a command was issued) then this command would not have been used as often although it could be assumed that actual navigation commands would have increased as there is a trade off in using speech for continuous or discrete movement. The most common navigation commands were the following: move, look, turn, walk, go; combined with supplementary commands such as forwards, backwards, right, or left. The least common navigation commands were side step, zoom in, shuffle, strafe. Whilst these commands were still quite general, they were used infrequently and therefore would not seem to be as intuitive as the more common commands. For a detailed breakdown of the commands by frequency please refer to Stedmon (2005).

In relation to object manipulation, the most common commands were the following: pick up, put down, and open; combined with supplementary commands such as screwdriver, screws, door. The least common commands for object manipulation were take out, hold, use. Such commands were relative terms to other items in the VE and participants did not use these as often in the HMI Trial. This relates to the idea that interaction devices should account for the type of manipulations a user has to perform and be designed, so that they adhere to natural mappings in how the device is manipulated, as well as permit movements that coincide with a user’s mental model of the type of movement required in a VE (Barfield et al., 1998).

From observations made during the user trials, the time taken to complete the task related more to navigation and orientation problems, rather than object manipulation. Participants often had difficulty orientating themselves to view specific aspects of the task or in navigating to a specific location in the VE in order to achieve the required viewpoint. The number of computer prompts observed between the two user trials was not significant but illustrated that participants used commands beyond the capabilities of the VE (e.g. a participant wanted to fly in the VE, another participant wished to zoom in on a particular detail). This kind of data illustrated the diversity of user requirements for completing the task and is valuable for developing future interaction capabilities in this, and other VEs.

The total number of commands that participants used to complete the task in both trials ranged from 89 to 415, illustrating a high degree of variability in task efficiency. The standard deviation for time to complete the task, total number of commands and total number of words was high and indicated the subjective nature by which participants developed task grammar and the impact this had on competing the task. The mean number of words per command had low standard deviations in both the HMI and HHI Trials which illustrated a high degree of consistency in the manner in which people interacted with a machine or another person.

4.2. Incorporating the findings into the RBI framework

By relating these findings back to the RBI framework (Jacob et al., 2008) it is possible to make some informed judgments and recommendations across the four themes:

- **Naïve physics**: from the most common commands, participants exhibited strong common sense knowledge of the world relating to physical movement and
interaction. VEs should therefore exploit these where possible to build intuitive contexts of use.

- **Body awareness and skills**: participants exhibited an awareness of and skills for controlling and coordinating themselves in the VE. This did not always translate into effective navigation but should be developed and supported in future VEs.

- **Environment awareness and skills**: participants negotiated the VE in a similar fashion to real world encounters but also exhibited desires for interaction in ways that are not possible in real world encounters. VEs offer an opportunity to expand environmental contexts and awareness.

- **Social awareness and skills**: participants did not interact with others directly in the VE but through instructions to control the interactions of other people or a machine. In doing this they exhibited different behaviours and strategies. Understanding this behaviour will lead to more effective use of speech input for HMI and HHI within future VR applications.

5. Conclusion

This paper has presented two independent user trials that were combined in one experimental design to investigate reality based interaction in a VR application. By allowing users to employ free speech it was possible to observe how users behave when restrictive vocabulary requirements are lifted. Overall, it would be difficult to develop a highly restrictive, command vocabulary for the task in this application due to the high degree of variance in the number of commands that participants used. However, this study supports the idea that a more open and general task grammar could be developed using the most popular commands that users might also find the most intuitive. Using a reality based interaction approach, the results from the user trials indicate that speech was considered to be beneficial for discrete tasks, object manipulation and short cuts whilst anecdotal evidence suggests that speech may not be well suited for specific actions such as continuous navigation. As a result of these findings, a combination of speech with other input devices might offer users a more flexible and integrated set of interaction tools for VR applications of the future.

Acknowledgement

This work was supported by IST Grant 2000-26089: VIEW of the Future. The authors are indebted to the anonymous reviewers of this paper who offered valuable and constructive feedback.

References


